7-30-2021

Literature-Assisted Validation of a Novel Causal Inference Graph in a Sparsely Sampled Multi-Regimen Exercise Data

Aditya Gupta
axg9642@rit.edu

Literature-Assisted Validation of a Novel Causal Inference Graph in a Sparsely Sampled Multi-Regimen Exercise Data

by

Aditya Gupta

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Bioinformatics

Thomas H. Gosnell School of Life Sciences
College of Science
Rochester Institute of Technology
Rochester, NY

July 30th, 2021

Contents

List of Figures & Tables

Rochester Institute of Technology
Thomas H. Gosnell School of Life Sciences
Bioinformatics Program

To: Head, Thomas H. Gosnell School of Life Sciences

The undersigned state that Aditya Gupta, a candidate for the Master of Science degree in Bioinformatics, has submitted his thesis and has satisfactorily defended it.

This completes the requirements for the Master of Science degree in Bioinformatics at Rochester Institute of Technology.

Thesis committee members:

Name	**Date**
Gary R. Skuse, Ph.D. Thesis Advisor	
Gordon Broderick, Ph.D.	
Matthew Morris, Ph.D.	

Feng Cui, Ph.D.
Director of Bioinformatics MS Program

475-4115 (voice)
fxcsbi@rit.edu

Abstract

Background. Causal mechanisms supporting the cardio-metabolic benefits of exercise can be identified for individuals who cannot exercise. With the use of appropriate causal discovery algorithms, the causal pathways can be found for even sparsely sampled data which will help direct drug discovery and pharmaceutical industries to create the appropriate drug to maintain muscles.

Objective. The purpose of this study was to infer novel causal source-target interactions active in sparsely sampled data and embed these in a broader causal network extracted from the literature to test their alignment with community-wide prior knowledge and their mechanistic validity in the context of regulatory feedback dynamics.

Methods. To this goal, emphasis was placed on the female STRRIDE1/PD dataset to see how the observed data predicts a Causal Directed Acyclic Graph (C-DAG). The analytes in the dataset with greater than 5 missing values were dropped from further analysis to retain a higher confidence among the graphs. The PC, named after its authors Peter and Clark, algorithm was executed for ten thousand iterations on randomly sampled columns of the modified dataset keeping intensity and amount constant as the first two columns to see their effect on the resultant DAG. Out of the 10,000 iterations, interactions that appeared more than 45%, 50%, 65%, 75% and 100% were observed. The interactions that appeared more than 50% of the times were then compared to the literature mined dataset using MedScan Natural Language Processing (NLP) techniques as a part of Pathway Studio.

Results. Full consensus across all sub-sampled networks produced 136 interactions that were fully conserved. Of these 136 interactions, 64 were resolved as direct causal interactions, 5 were not direct causal interactions and 67 could only be described as associative. It was found that about 17% of the interactions were recovered from the text mining of the 285 peer-reviewed journals from a total of 64 that were predicted at a 50% consensus. Out of these 11, 4 were completely recovered whereas 7 were only partially recovered. A completely recovered interaction was LDL → ApoB and a partially recovered interaction was HDL → insulin sensitivity.

Conclusion. Only 17% of the predicted interactions were found through literature mining, remaining 83% were a mix of novel interactions and self-interactions that need to be worked on further. Of the remaining interactions, 53 remain novel and give insight into how different clinical parameters interact with the cholesterol molecules, biological markers and how they interact with each other.

Introduction

Graphical Causal models are gaining importance in the area of muscular physiology. However, to identify mechanisms which support the benefits of cardio-metabolic exercise, there are experiments performed separately for some biomolecules, but no single pathway has been published that could give us answers about how to maintain muscles more efficiently. Our research explored new and unusual causal relationships among the various study parameters observed that might prove useful to physiology as well as to personalized medicine. The expectation is that the results of this work shall help us treat and maintain muscular functions in comatose patients and in those professions with minute exercise regimes as well as in conditions of microgravity related to space travel because individuals in these scenarios get very little to no exercise to be able to maintain healthy muscular physiology. Lastly, the interaction maps shall also provide insight into improved recovery from cardiac events and heart muscle damage. In this use of causal discovery pathways, we focus on improving our understanding of muscular pathways from the Studies Targeting Risk Reduction Interventions through Defined Exercise (STRRIDE) study performed by the Duke University's Molecular Physiology Institute (DMPI) (Johnson et al. 2019). Interestingly, the study revealed that the moderate intensity exercise regimen proved to reduce the greatest fasting insulin measure in the 10 years after the original study was performed in the early 21st century (Kraus et al. 2001).

An ever-growing database of peer-reviewed journals have and continue reporting and improving new and existing biological mechanisms that tie together interactions among biomolecules found in the human body. Their interactions are not always straightforward enough to catch the eyes of even the researchers who are working on them. Some of these issues were explored in a study by Ferreiro and coworkers (Ferreiro, Komives, and Wolynes 2014). Literature sources prove to be very useful for identifying relationships between two or three biomolecules due to the less complex networks involved, however, to develop an interaction map, they do not include enough data (de Las Rivas and Fontanillo 2010). Rarely do they report a complete cyclic relationship with sparsely sampled data. Our research addresses the problem of causal inference structure with a target molecule of interest using previously observed data. Through this study, we shift our focus to the discovery of local direct causes or direct effects of the target against a significant number of other variables. Knowing about the direct causes and

effects, we can predict mechanisms that can prove helpful for the drug industry to develop specific drugs targeting the mechanism in the host's body that proves to be irregular.

Traditionally, causal models were discovered by randomized trials and clearly laid-out interventions. Though providing the highest level of confidence, this can be very inefficient, costly and sometimes impossible. Hence, using observational data to predict a cause-effect relationship from this data using Bayesian networks for causal discovery has gained momentum. The observational data can be collected without controlling factors that might be hypothesized to affect the system in question (Rathnam, Lee, and Jiang 2017). Time series data is fairly popular when it comes to using some of the advanced causal discovery algorithms (Hyttinen et al. 2016). However, since the STRRIDE data includes only two time points, i.e., pre- and post-intervention, using a simpler algorithm for causal discovery that does not build on rate equation formulations, seemed more plausible. Another reason for this is that the STRRIDE data cannot be simply viewed as a large time-series dataset with missing time steps. This would create an issue with the reliability of the current data which was not the goal of this study.

The purpose of this study was to infer novel causal source-target interactions active in sparsely sampled data and embed these in a broader causal network extracted from the literature to test their alignment with community-wide prior knowledge and their mechanistic validity in the context of regulatory feedback dynamics. To achieve this, the data was verified for consistency. Once the consistency thresholds were defined, data was stratified to fit an arbitrary schema that would ensure homogenous data that would be an input to a causal discovery algorithm of choice. The output of this algorithm would be a directional graph that can be compared and analyzed with literature mined graphs for similarities. This would ensure the research would be directed towards a number of undiscovered edges that can be verified with further experimentation.

Approach

The discovery of native causal relationships is very vital as it plays a central function in causal discovery and classification of interactions between biomolecules and their governing mechanisms (Salon, Lodowski, and Palczewski 2011; Subramaniam et al. 2011; Cyr and Domann 2011). The structure of interaction networks is highly scalable across levels of biological complexity thanks to their scalable edge density distributions, e.g. individual

biomolecules involved in a certain pathway can be upregulated or downregulated in an experimental setup to find the appropriate mechanism of action to fight a disease and create a personalized drug. The Peter-Clark (PC) algorithm explores one such opportunity using Bayesian conditional independencies among the different individual markers in a dataset (Spirtes and Glymour 1991; Spirtes, Glymour, and Scheines 2000; Kalisch and Bühlmann 2007). The PC algorithm (Spirtes, Glymour, and Scheines 2000) provides a computationally efficient and reliable output given the faithfulness of conditional independence among the different variables, i.e. the variables can have conditional independencies among them that appear at higher order (involving more than two variables at once) independence relations.

The predictions of these discovery pathways generally bridge causation with predictivity, giving us more information about those edges and their inherent interactions. Determining the native causal inference map gives us more details about the natural and predicted pathways (using computer programs) which in turn helps us decide the best interventions to help us achieve the desired behaviors from the model organism, although, certain assumptions are made based on the nature of input data. One such assumption states "A variable X is independent of every other variable (except X's effects) conditional on all of its direct causes" (Scheines 1997) which tells us that each variable in the PC algorithm is treated independently whereas in reality that might not always be the case. Generally, due to an overlap in the functions of certain protein markers, i.e. one protein marker affects the other protein marker in a biological pathway. For example, point mutations in the Adenomatous Polyposis Coli (APC) Mutation Cluster Region (MCR) leads to disabling the *wnt* signaling pathways depending on the effect of the mutation (Minde et al. 2013). Understanding the local pathways shall ultimately help us understand the role of every edge on our map on a global scale (Silverstein et al. 2000; Nikolay et al. 2017). This can be translated into novel algorithms that prove to be more time-efficient and are flexible enough to suit the needs of our goals.

There are certain algorithms for the statistical inference of causal relationships that are known to us in this field. We explore the PC algorithm (Spirtes, Glymour, and Scheines 2000) to find all interactions predicted in our dataset based on only the values of the experiments and draw conclusion from a biologist's perspective using natural language processing (Novichkova, Egorov, and Daraselia 2003) to find the known relations and inform us about new ones that can be verified with targeted protein interaction studies. Although most biological pathways are

cyclic in nature, we try to find acyclic pathways because cyclic pathways have no start or end node. In a cyclic pathway, interactions can be represented by multiple acyclic pathways (Strobl 2019).

In this work, we applied the PC algorithm to the STRRIDE data to provide insight into interactions among the sparsely sampled analytes that can be verified by the resampling of analytes as well as with the literature mined sources. The Bayesian conditional inference rules help predict these interactions that are a mix of literature mined interactions and some novel ones to be studied through further experimentation. The PC algorithm provides us with one such resource to predict a large number of interactions based on raw data from a large-scale experiment such as the STRRIDE study.

Methods

Subjects, study characteristics and data

The participants recruited into the 3 STRRIDE studies were from North Carolina communities near the Duke University. They were 40-65 years old with a sedentary lifestyle. The exercise regimen was practiced over a period of six months and all analytes were measured before the intervention and after the six-month duration of the intervention. The experimental protocol included varying levels of the amount of exercise as well as varying levels of the intensity of workout. For the amount of exercise, the prescription varies from 14kcal per kg body weight for the low amount to 23kcal per kg body weight for the high amount. The intensity of exercise varied from 65-80% peak oxygen consumption for the vigorous intensity and 40-55% peak oxygen consumption for the moderate intensity of exercise.

The edges for the causal interaction pathway include both clinical and physiological parameters. These were observed in a pre-intervention state as well as in a six-month post-intervention state from 590 patients (randomly distributed among men and women) with varying levels of adherence to the proposed regimens. A comprehensive list of these analytes is listed in Supplementary Table 1.

The data were retrieved from the Duke Molecular Physiology Institute (DMPI), Duke University in raw format which comprised one large dataset from 317 participants (with >75%

adherence to their respective exercise protocol) from all three STRRIDEs pooled into it with the levels of interventions being Low-Amount Moderate Intensity, Low-Amount Vigorous Intensity, High-Amount Vigorous Intensity [STRRIDE 1]; Aerobic Training, Resistance Training, Aerobic plus Resistance Training, High Amount Aerobic Training (n=~10) [STRRIDE 2], Diabetes Prevention Program, Low-Amount Moderate Intensity, High-Amount Moderate Intensity, High-Amount Vigorous Intensity [STRRIDE PD]. An observation was noted for all the characteristics of the data. The types of analytes fell into these categories:-

1. Protein – protein weight detection
2. Small Molecules – macromolecules that were measured
3. Clinical Parameters – analytes were measured using clinical techniques and do not directly specify a biomolecule.
4. Independent – Amount and Intensity of interventions
5. Unknown – analytes whose names were not found in the key

Pre-Processing

We found two options for combining data subsets in order to explore all possible permutations of Intensity and Amount of exercises to work with, so the data were grouped into. {1} Low-Amount Moderate Intensity, Low-Amount Vigorous Intensity, High-Amount Vigorous Intensity data retrieved from STRRIDE 1 study and High-Amount Moderate Intensity data from STRRIDE PD study. {2} Low-Amount Vigorous Intensity data from STRRIDE 1 study and Low-Amount Moderate Intensity, High-Amount Moderate Intensity, High-Amount Vigorous Intensity data from the STRRIDE PD study. This was done by matching rows among the studies with the respective Intensity and Amount of the intervention.

Statistical Analysis

An ANOVA was performed with the interaction of amount and intensity as the independent variables to find if they had an effect on the delta value of the pre-intervention and post intervention analytes within four interventions, namely; Low-Amount Moderate Intensity, Low-Amount Vigorous Intensity, High-Amount Vigorous Intensity (STRRIDE 1); High-Amount Moderate Intensity (STRRIDE PD) as well as on the basis of sex. The selection of significantly varying analytes was selected on the basis of p-score being less than 0.05. The p-scores were

then corrected using Benjamini-Hochberg (BH) correction and the corresponding corrected p-value (q-value) cutoff being 0.05, i.e. q <= 0.05. Box and whiskers were plotted for the q-values using ggplot2 package whereas the raw p-values from the ANOVA were tabulated.

Missing Data Reporting

There were on an average 30.3% (Min:2.6% Max:80.3%) missing values (indicated by NA/NaN) found in the data. These were summarized and subjected to having a maximum of five NA values per each of the four interventions per analyte to help decide which analytes to use for the causal inference network discovery.

Network Assembly

Empirical networks

The pcalg (Kalisch et al. 2012; Hauser and Bühlmann 2012) package in R was used to find the initial Directed Acyclic Graph (DAG). The data was extracted for the significantly changing analytes (p-value <= 0.05) from the two-way ANOVA on the STRRIDE study using the pre and post intervention values of analytes. The PC algorithm was run on this dataset multiple times to get correct directions of interactions that were biologically viable.

A delta change metric was used to make time implicit and model the approximate rate of change. The pre-intervention values were subtracted from post intervention values from all the analytes that were to be used further. Another metric used to remove the bias was a fold change metric which was the delta change metric divided by pre-intervention values. The delta change metric was given preference because it was found that the fold change created errors in dataset due to division by zero errors where delta change did not. Note, that an existing NA value in either the pre-intervention or the post intervention shall make the resultant value as NA as well.

To compensate for bias introduced by the ordering of input variables, PC algorithm was run through 10,000 iterations on the input dataset where each column of the dataset represented an analyte that would impact each of the 10,000 output graphs. The algorithm generated a list of directed and undirected interactions between nodes for each of the iterations which were stored in Rdata format. The bidirectionality was seen due to PC not being able to identify a directed edge (a direct causal relation) between two nodes (analytes). To address this concern, the bi-

directional edges were separated from directed edges and set aside for comparison with literature informed networks. This was done because the algorithm would predict with certainty if there was a causal relationship between the two analytes as a unidirectional edge. From the output unidirectional graphs, graphs where edges that were found in 45%, 50%, 65%, 75% or 100% of the iterations were retained and the causal inference graphs were plotted accordingly to the retaining threshold. A simple graphing tool (yED and iGraph Package, R) was utilized to view the graphs that would be generated in all these cases. To analyze the characteristics of these graphs, Cytoscape (Paul Shannon et al. 1971) was used. Cytoscape with the *NetworkAnalyzer* tool provides shortest paths, centrality measurements, clustering coefficients for both directed and undirected graphs.

Literature informed networks

As a measure of validation against the published literature, the graph edges were recorded and compared with the Natural Language Processing (NLP) output from the Pathway Studio tool. The NLP algorithm was used to data mine the interactions from the Elsevier database of journals. The NLP algorithm is derived from MedScan, which uses PubMed abstracts and full-text articles from the PubMed database. MedScan NLP pulls out biological network information such as cellular processes, clinical parameters, complexes and biomolecules such as proteins as well as other small biomolecules such as high-density lipoproteins. The analytes from the study were entered into the Pathway Studio and it provided literature mined interactions among the analytes as entities and also returned a KEGG ID for Pathways wherever applicable, a direction of relation provided, types of source and target analytes as well as the total number of references and specific sentences containing the relation which it found in the database. The total number of references for the interactions found among these analytes was 7,015 among which on an average, 37 references were found for each interaction.

This enabled us to label interactions as either *Complete*, *Partial* or *None* (not existent) in comparison to the PC results. The interactions were labelled as *Complete* when the NLP produced a result with the exact direction of edge found in the PC algorithm output. A *Partial* label was awarded to edges that had either the directions of source to target swapped or one of nodes were a more resolved molecule than Pathway Studio allowed us to. *None* was assigned to interactions that the NLP tool failed to find. This was done by manually comparing the

spreadsheets produced by NLP and comparing them to a source-target list of interactions extracted from the predicted directed acyclic graph.

Results

Data Pre-processing

Statistical Analysis

Table 1 shows the result of a two-way analysis of variance (ANOVA) applied to the normalized dataset based on levels of two variables, namely amount and intensity. The significantly changing variables can be used to predict the causal inference graph due to very low p-values, however, on applying the method to male and female datasets, it was observed that very high missing value analytes disappeared. On further applying the 2-way ANOVA for option 1 resulted in only 12 variables showing a significant difference in their mean expression. There were NAs found in the Option 1 (Low-Amount Moderate Intensity, Low-Amount Vigorous Intensity, High-Amount Vigorous Intensity data retrieved from STRRIDE 1 study and High-Amount Moderate Intensity data from STRRIDE PD study) data due to one of the pre- and post-values missing from the original dataset (Supplementary Figure 2 & Supplementary Table 2). The ANOVA for Option 2 (Low-Amount Vigorous Intensity data from STRRIDE 1 study and Low-Amount Moderate Intensity, High-Amount Moderate Intensity, High-Amount Vigorous Intensity data from the STRRIDE PD study) revealed that 12 variables changed significantly in their delta change values over the duration of the intervention. A substantial number of missing values were noted in this stratification of the whole dataset so Option 1 was chosen for further analysis.

Table 1: Two-way ANOVA on the complete dataset with combinations of moderate vs vigorous intensity and low vs high amount. The missing values are reported in the last column.

Complete Data	Df1	Sum Sq1	Mean Sq1	F value1	Pr(>F)1	Missing Values
age	3	5.61E-01	1.87E-01	1.55E+01	1.87E-09	2
weight_kg	3	7.38E+01	2.46E+01	3.82E+00	1.03E-02	12
waist_circum_cm	3	1.17E+02	3.89E+01	4.22E+00	6.15E-03	52
avo2	3	1.70E+02	5.67E+01	9.52E+00	5.14E-06	32

rvo2	3	5.75E+01	1.92E+01	2.96E+00	3.27E-02	32
matsuda	3	3.68E+02	1.23E+02	1.53E+01	2.85E-09	21
sbp	3	2.01E+05	6.69E+04	4.43E+01	1.19E-23	26
dbp	3	1.91E+05	6.36E+04	4.32E+01	5.16E-23	36
albumin	3	1.46E+00	4.86E-01	3.45E+00	1.77E-02	130
cmv	3	7.50E-01	2.50E-01	3.18E+00	2.50E-02	130
h6p	3	2.32E+00	7.72E-01	3.04E+00	2.93E-02	37
hsa_mir_223_3p	3	2.91E+01	9.70E+00	3.20E+00	2.47E-02	150
Males	**Df1**	**Sum Sq1**	**Mean Sq1**	**F value1**	**Pr(>F)1**	**Missing Values**
age	3	2.22E-01	7.40E-02	5.60E+00	1.15E-03	2
avo2	3	9.38E+01	3.13E+01	4.00E+00	9.17E-03	20
lbm	3	9.64E+00	3.21E+00	3.02E+00	3.16E-02	8
matsuda	3	2.60E+02	8.65E+01	9.03E+00	1.59E-05	8
sbp	3	9.44E+04	3.15E+04	2.15E+01	1.42E-11	10
dbp	3	9.53E+04	3.18E+04	2.29E+01	3.72E-12	13
bun	3	1.13E+02	3.78E+01	2.98E+00	3.61E-02	73
gsp	3	3.44E+04	1.15E+04	3.20E+00	2.55E-02	21
cmv	3	1.54E+00	5.15E-01	4.09E+00	9.33E-03	73
infx	3	3.64E+02	1.21E+02	3.61E+00	1.52E-02	21
Females	**Df1**	**Sum Sq1**	**Mean Sq1**	**F value1**	**Pr(>F)1**	**Missing Values**
age	3	4.14E-01	1.38E-01	1.29E+01	1.28E-07	2
avo2	3	6.66E+01	2.22E+01	5.30E+00	1.68E-03	14
rvo2	3	6.32E+01	2.11E+01	5.04E+00	2.36E-03	14
matsuda	3	1.37E+02	4.57E+01	7.75E+00	7.63E-05	15
sbp	3	1.02E+05	3.41E+04	2.26E+01	4.48E-12	18
dbp	3	8.62E+04	2.87E+04	1.88E+01	2.73E-10	25
albumin	3	2.00E+00	6.65E-01	5.53E+00	1.45E-03	59
crp	3	1.62E+04	5.38E+03	5.03E+00	2.69E-03	59
apob	3	1.45E+03	4.84E+02	3.76E+00	1.23E-02	18
h6p	3	3.71E+00	1.24E+00	3.96E+00	9.54E-03	18
nldlc	3	2.56E+03	8.52E+02	3.62E+00	1.47E-02	18
totchol	3	4.30E+03	1.43E+03	3.94E+00	9.78E-03	18
hsa_mir_133a_3p	3	1.32E+02	4.41E+01	3.03E+00	3.31E-02	69
mir_374b	3	4.06E+01	1.35E+01	3.02E+00	3.36E-02	69

The ANOVA revealed certain trends in the data. Age was seen in all raw p-score sorted analytes as was expected because age increased at a fixed rate over the duration of six months from pre to post intervention. The BH correction removed analytes that were potentially false positives, hence, the most significantly differing analytes between pre- and post-intervention were extracted. Figure 1A-C showed that the Matsuda index, systolic and diastolic blood

pressures significantly changed over the period. However, the direction of change varied significantly due to multiple lines connecting the pre- and post-intervention analytes increasing and decreasing at the same time. The medians of these analytes, however, changed as expected, since the Matsuda index increased slightly, and the blood pressure readings dropped significantly. The analytes in men also differed from women in the BH corrected graphs. Female subjects, in addition to significantly changing analytes from the whole dataset and men, showed changes in absolute and relative O_2 measures as well as albumin and c-reactive protein values also changed. There was a slight decrease in the medians of the avo2 and rvo2 whereas the medians for albumin and c-reactive protein did not show a visible change. A comprehensive list of missing data can be seen from supplementary figure 2.

Figure 1 A

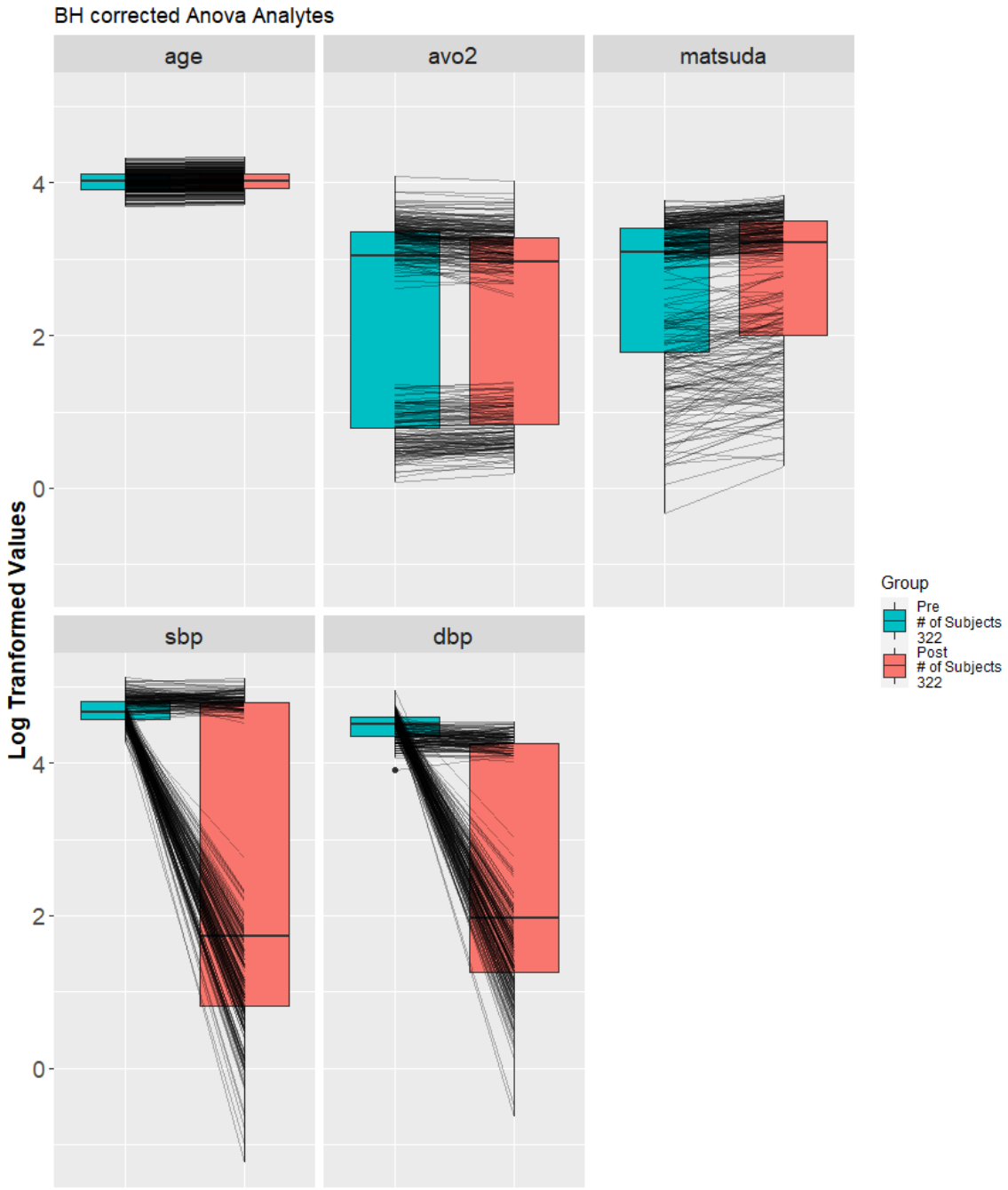

BH corrected Anova Analytes
age
avo2
matsuda
sbp
dbp
Log Tranformed Values
4
2
0
4
2
0
Group
Pre
# of Subjects
322
Post
# of Subjects
322

Figure 1 B

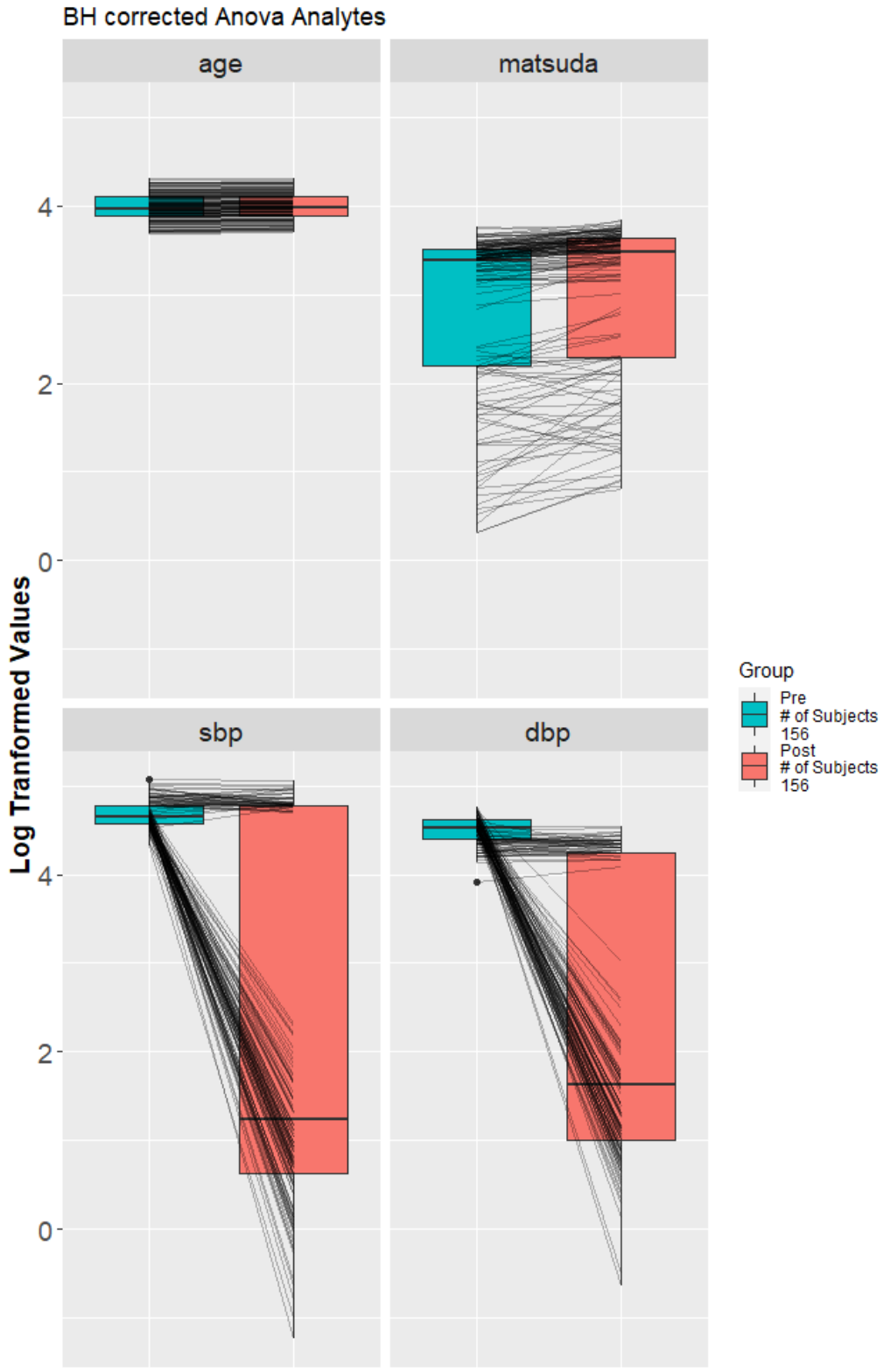

BH corrected Anova Analytes
age
matsuda
sbp
dbp
Log Tranformed Values
4
2
0
Group
Pre
# of Subjects
156
Post
# of Subjects
156

Figure 1 C

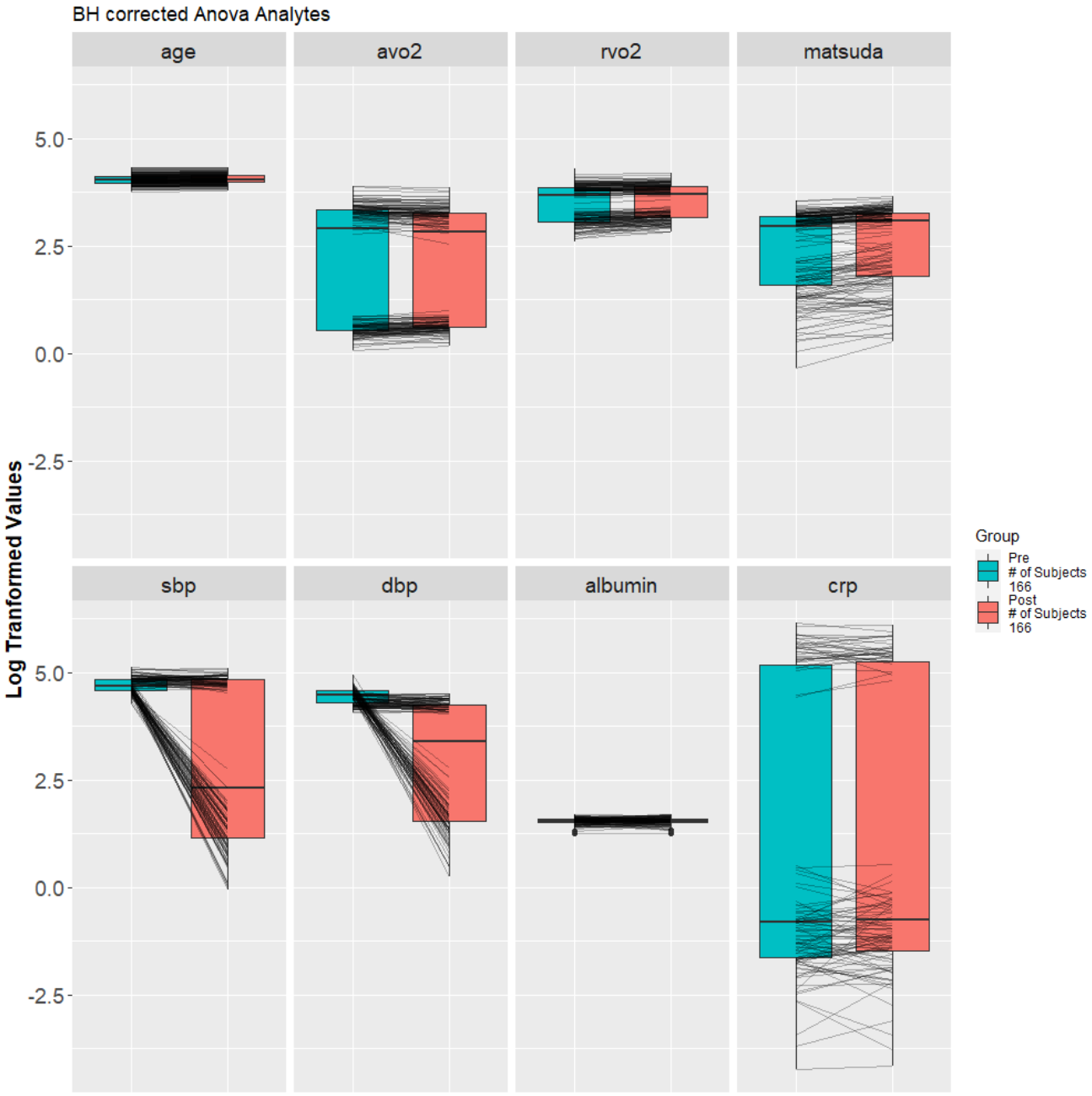


Figure 1: The significantly varying analytes from the STRRIDE studies using raw p-scores (left) and Benjamini-Hochberg corrected scores (right) among the males and females. (A) Log Transformed two-way ANOVA results for analytes for the whole dataset with pre vs post exercise intervention. (B) Log Transformed two-way ANOVA results for analytes for the male subjects with pre vs post exercise intervention. (C) Log Transformed two-way ANOVA results for analytes for the female subjects with pre vs post exercise interventions. The Pre-intervention analytes (turquoise) and post int. analytes (light red) values are connected through black lines for each value. The total number of subjects were All Data: 322; Male: 156; Female: 166.

Inferring Causal Interactions

Causal relationships between the analytes were used to predict the presence of a causal arrow in the resulting graph where a single-headed arrow showed clear X is a cause of Y (X→Y), a double-headed arrow representing an unclear relationship between X and Y (X↔Y). The direct causes and effects are predicted using PC algorithm. It works on providing local causal interactions based on one variable and helps with the directionality of the nodes in our resulting graph.

A cutoff of ≤5 NA was chosen for all columns with missing values. This ensured a completeness of the analytes so that a missing data column does not influence the output of the DAG. Once the higher NA count columns were removed from our data, we resampled our columns keeping our intensity and amount columns as the first 2 variables since they were meant to be source nodes in all cases, i.e. they drive the changes in analytes that we wish to see. The PC algorithm was then run for 10,000 iterations to see the positional effect of the order of input variables to the algorithm and to counter a possible bias introduced by the naïve PC algorithm. It was noted that positionality affected the output of the algorithm significantly. Hence, running a large number such as 10,000 iterations of PC on the Option1 Female stratified data was entered as the input to PC.

Different levels of consensus were observed to get a better understanding of which interactions were highly supported by the PC algorithm analyses. From Figure 2A-D, at 100% consensus, 23 interaction edges were observed in the graph. There were a large number of orphan nodes observed and 11 sub-networks were observed. As the consensus threshold was relaxed by dropping the value of threshold at 75%, 25 interactions were observed which was due to addition of 2 new interactions with less unanimously agreed upon edges but the sub-networks were still 11. At a threshold of 65%, 6 more interactions appeared making it 31 total interactions and dropped the sub-networks number to 10. At the threshold of 50%, a large singular graph with 63 interactions and an island graph of one interaction was observed. We found that forcing the conservation of an edge in over 50% of the networks created a larger number of disjointed acyclic graphs, most of which were essentially subnetworks of the larger one. Our motivation was to explore the full structure of the more connected graph even if it suffered a higher degree of uncertainty. From Figure 2E, dropping the threshold to 45% retention, the edges started to

show bidirectionality which was to be avoided. This was essential as only the directed edges from the out of 10,000 iterations of PC were used for further study. The bidirectional edges appeared due to some interactions having their cause-and-effect nodes flipped due to bias in position of analytes in the input of the PC algorithm. As a result, only interactions appearing 50% or more times were further studied as shown in Figure 2D. The edges shown in this graph included interactions between weight and fat mass / lean body mass; cholesterol and apolipoprotein B.

Figure 2 A

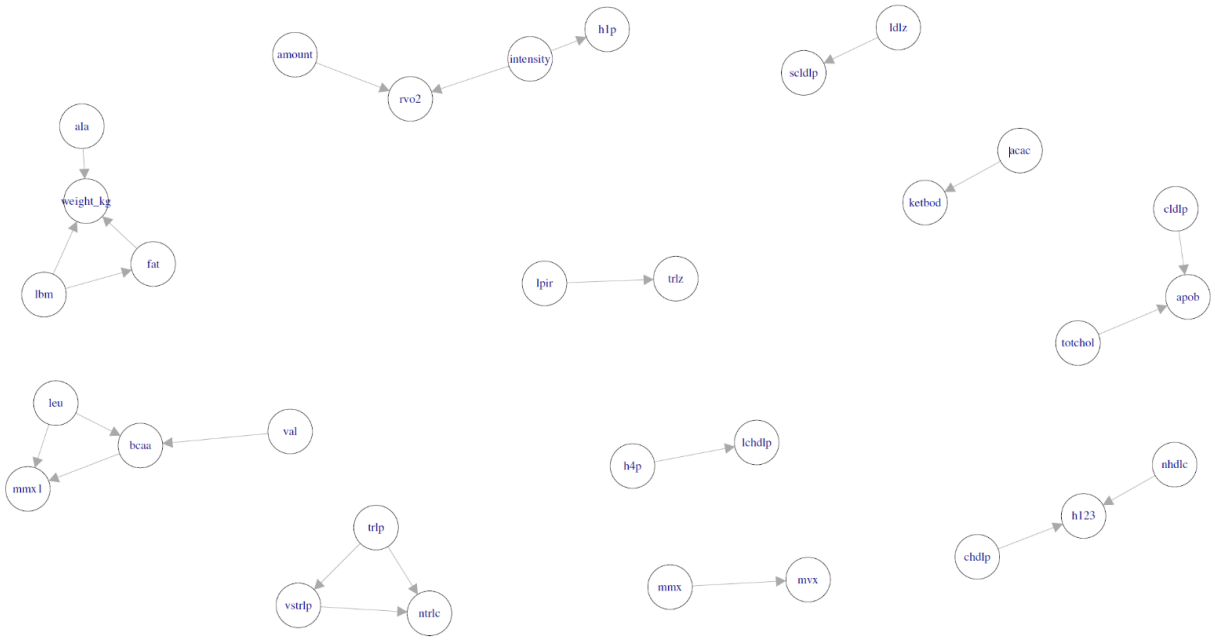


Figure 2 B

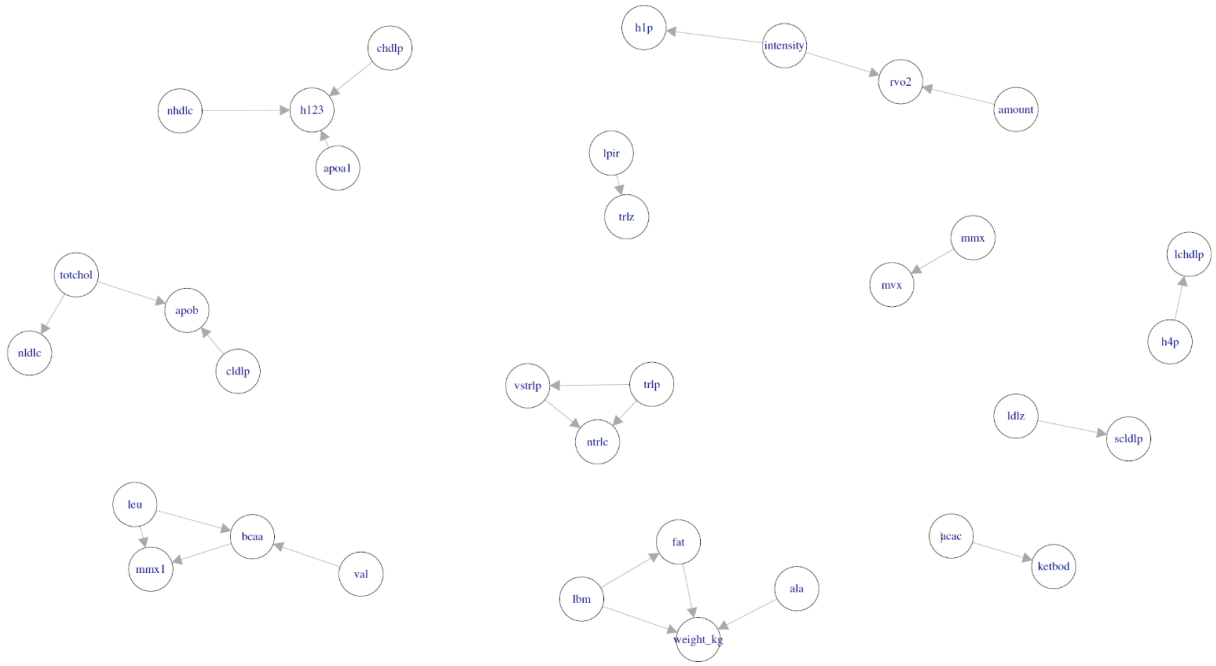


Figure 2 C

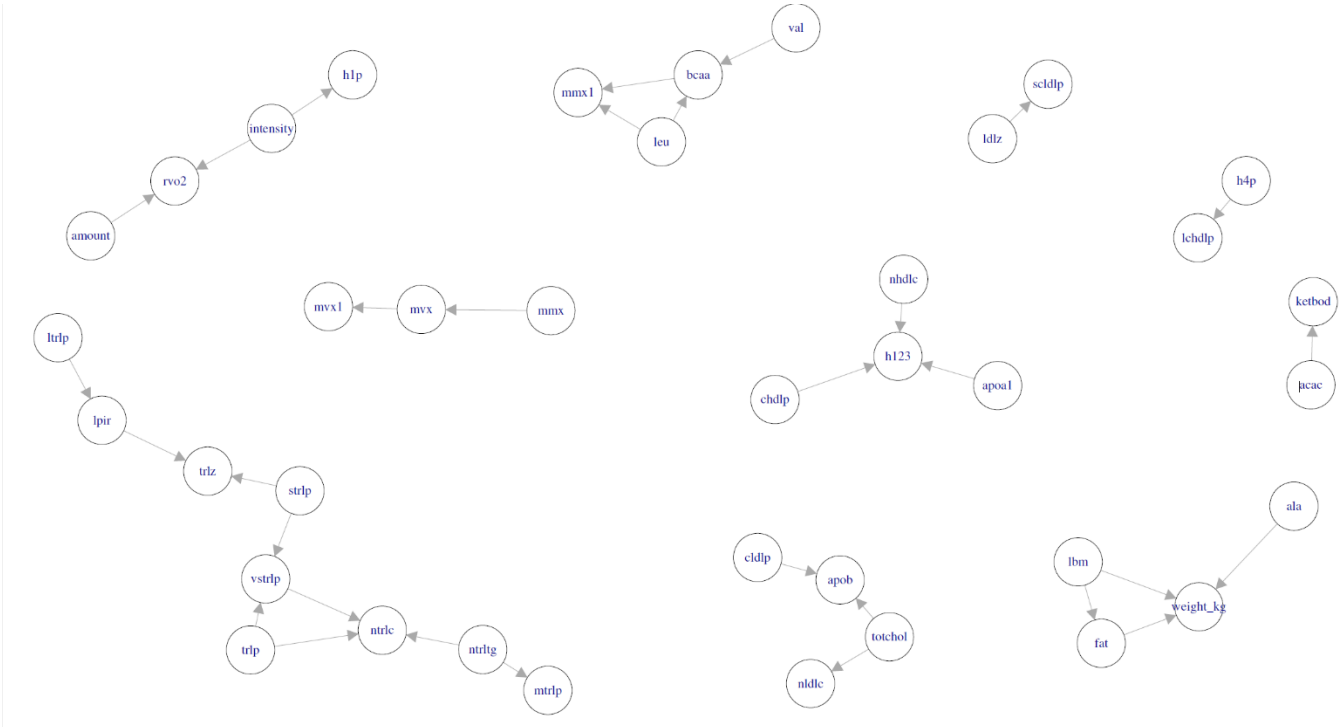


Figure 2 D

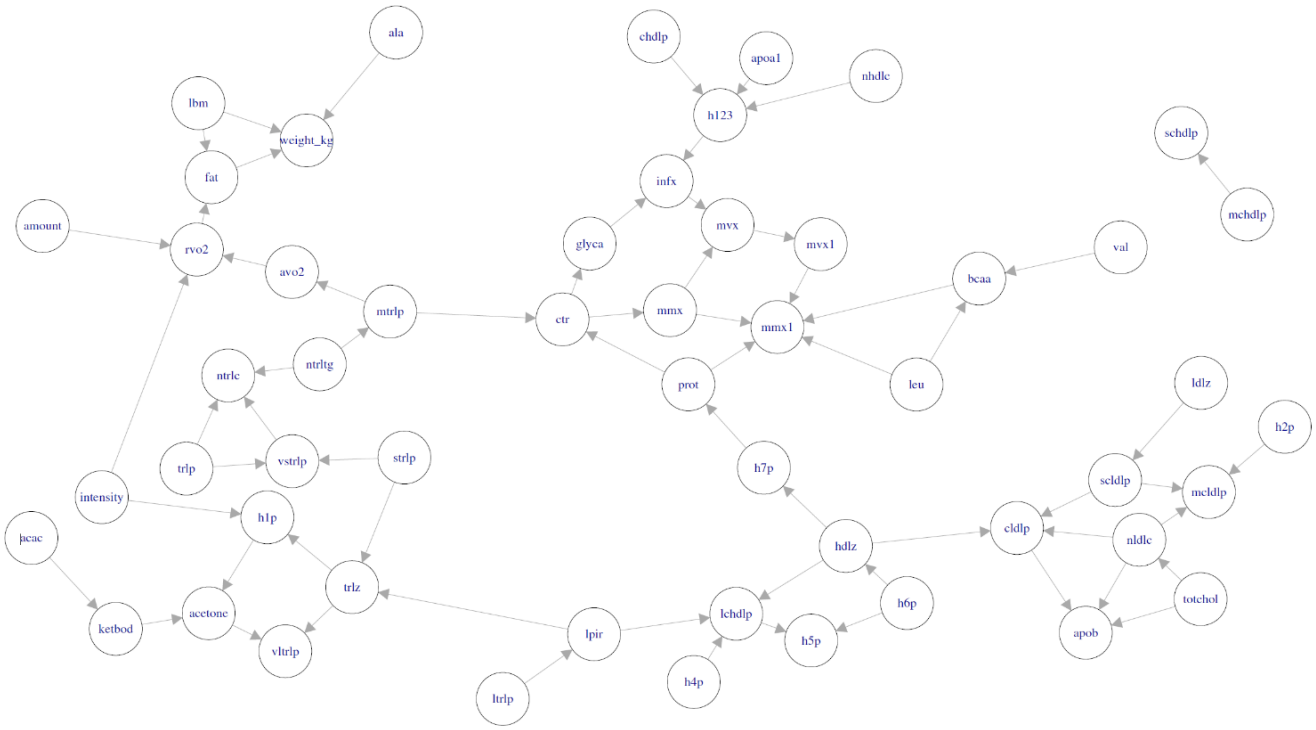


Figure 2 E

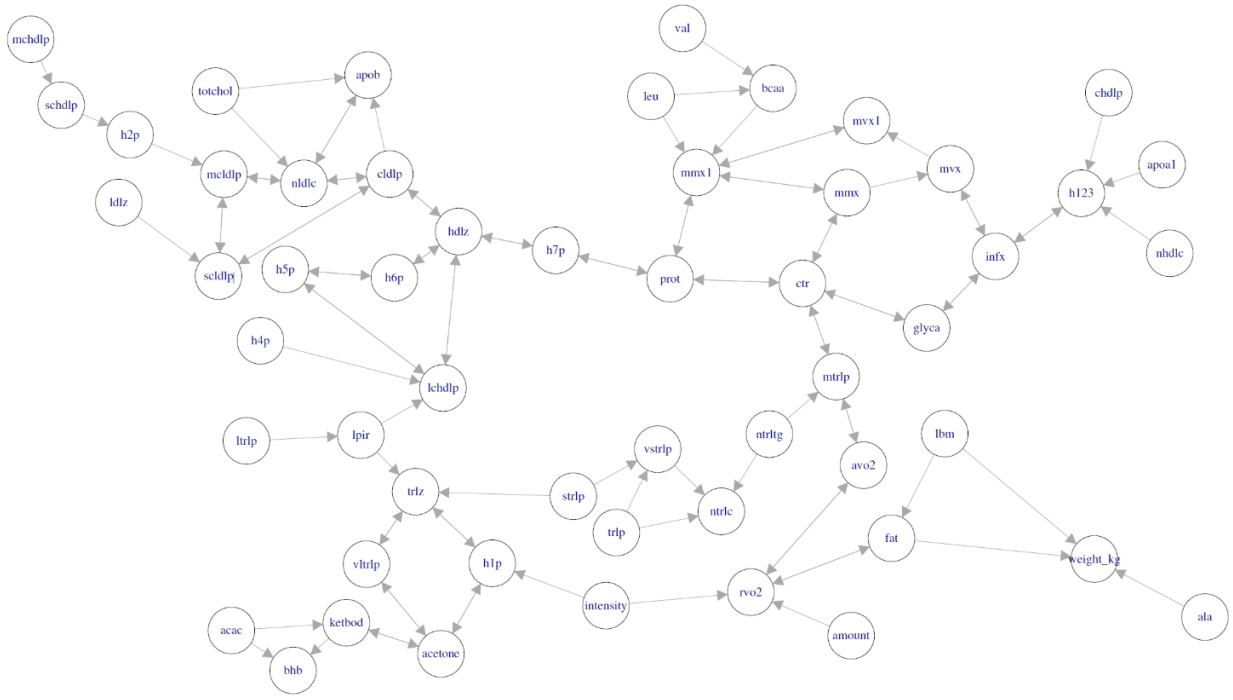


Figure 2: The graphs for option 1 10,000 iterations of PC algorithm. The edges that appear in (A) 100%, (B) 75%, (C) 65%, (D) 50% and (E) 45% of the total number of iterations are represented in the above figure. A node appearing in 65% graph means it appeared at least 6,500 times in a 10,000 runs of PC algorithm. All interactions provided were unidirectional until consensus of 45% was reached.

Figure 3 shows all interactions that were found in PC run from above with a 50% threshold line (black). There were only 5 undirected edges that appeared more than 50% consensus. The Directed Edges (blue) were used to make a large unidirectional graph shown in Figure 4. These were 64 interactions with all the analytes stratified into their classes and edges with weights according to their consensus level. Cytoscape network analysis revealed that the average number of neighbors was 2.415, network diameter was 7 which tells that the maximum length of the shortest path between two nodes. The characteristic path length was 2.361 which tells us the average shortest path length between two connected nodes. The network radius is 1 which the shortest length between any two nodes in the network. The network is sparsely populated as the density is 0.023. The clustering coefficient is 0.065 which is the ratio of number of edges in the neighbors of nodes to maximum edge count that is possible between nodes. The low value suggests that the graph is not densely packed.

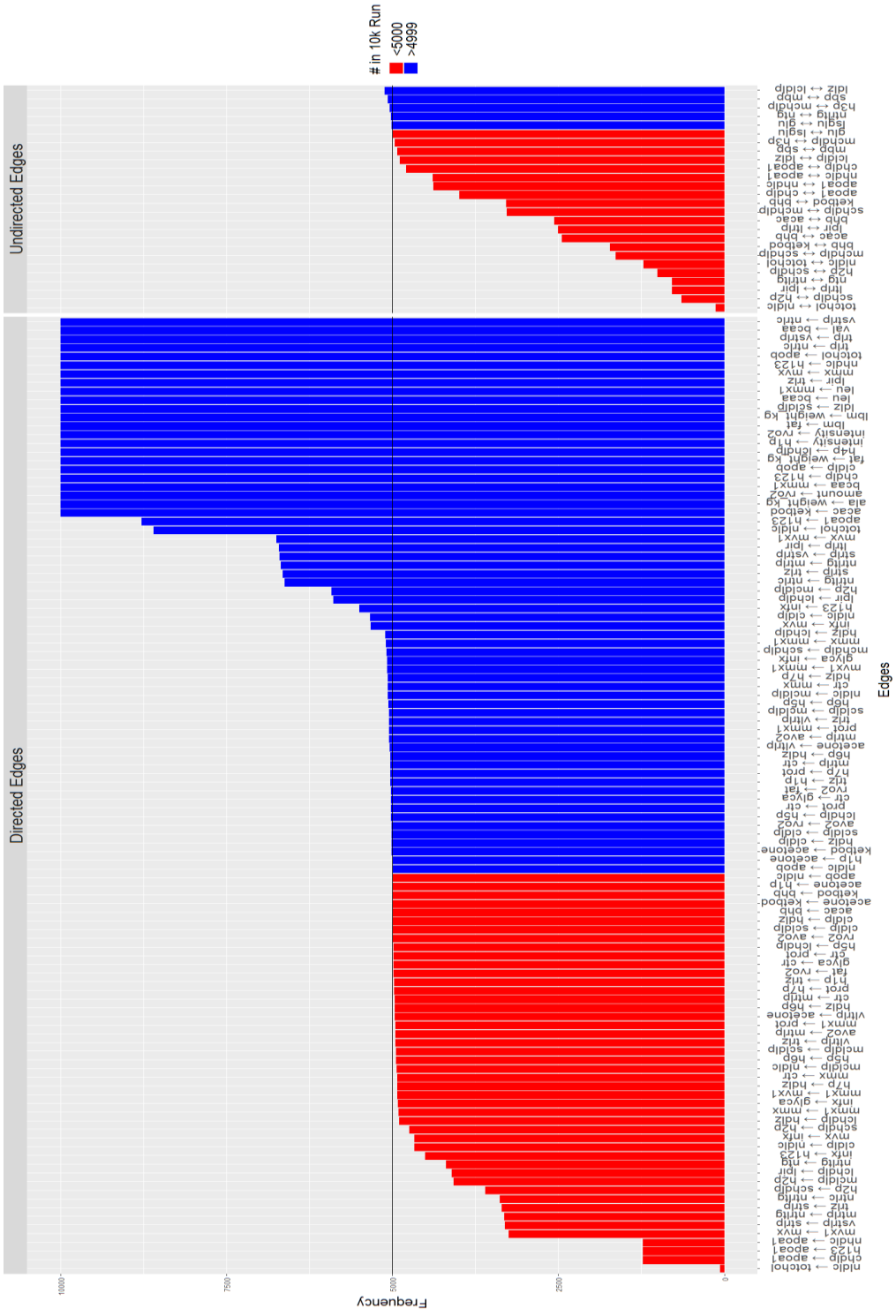


Figure 3: Number of the edges found in a 10,000 iterations of PC algorithm ran on option 1 data. The total number of directed and undirected edges is 136 where 110 directed edges and 26 undirected edges were present. The red bars represent the edges that were occurring less than 5,000 times and the blue bars represent edges occurring more than 4,999 times. There were 64 blue directed and 5 blue undirected edges appeared in at least half of the total runs.

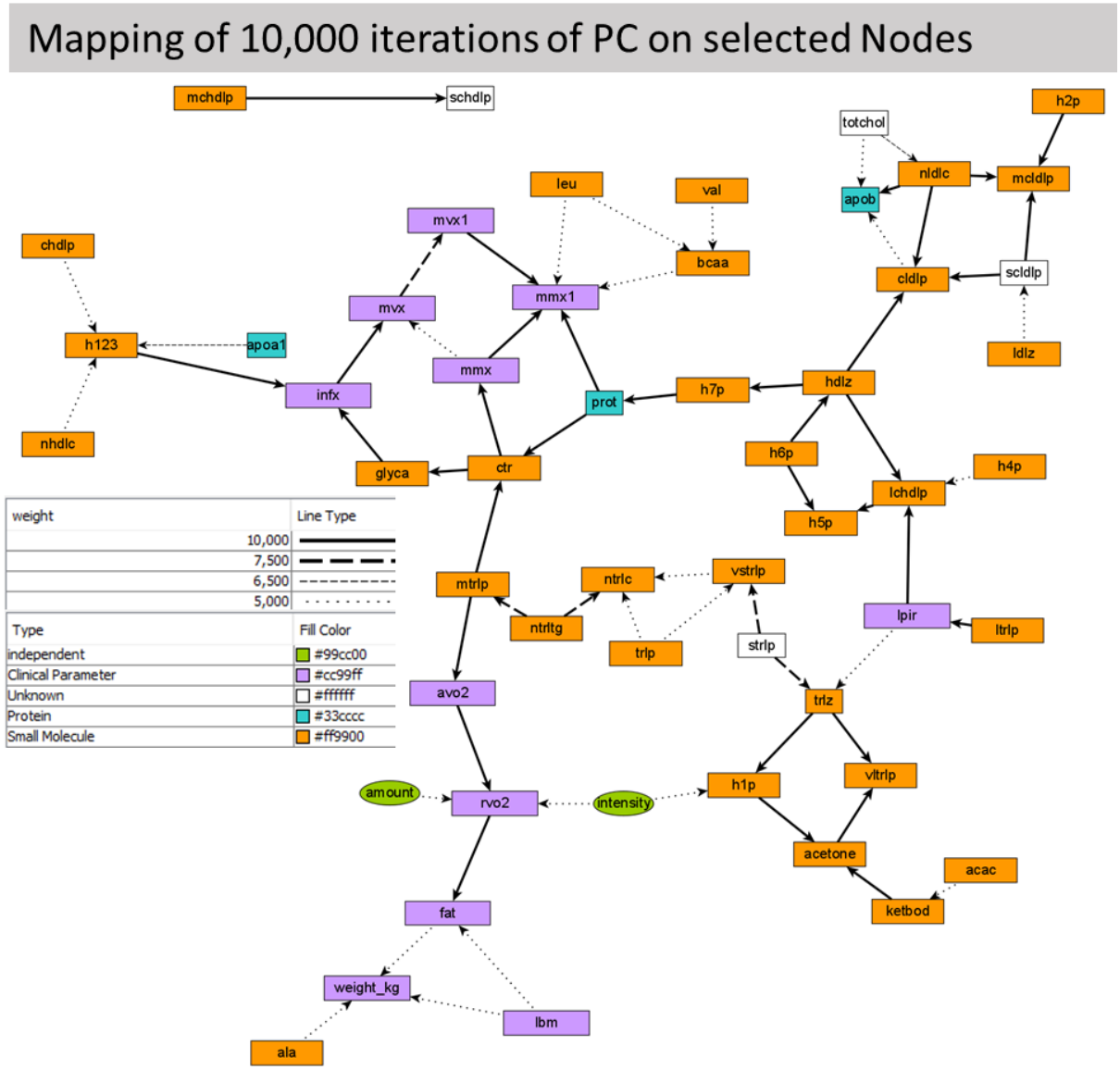


Figure 1: Figure 4: Option 1 DAG for Females. Each node stands for one of the variables from the experiment from STRRIDEs 1 and PD. The disjoint nodes are placed on the upper-right corner of the graph. Single-headed arrows show a causal relationship, for example, absolute O2 is a cause of relative measure of O2. The double-sided arrows represent an unclear causal relationship. This is a plot for 50% retention of nodes, i.e. nodes that appear in at least 5,000 occurrences out of a total 10,000 PC simulations.

Validation with a literature informed network

The NLP output was either a direct or indirect regulation retrieved from the peer-reviewed journals. It also gave information about the effect of the regulation, which was either positive, negative or unknown. There was a total of 186 interactions that were listed in the NLP output that were verified by 7,015 literature references. Of the 186 relations provided, 40

relations were direct and 146 were indirect. 44 of these relations were source analyte positively affecting the target, 64 had their source analyte negatively impacting the target analyte and 78 had an unknown effect on their targets. The types of analytes belonged to one of the same criteria as before:

Protein / Clinical Parameter / Small Molecule / Complex.

On comparison of the 50% retention graph with the NLP supported interactions, it was noticed that 17.82% of the 64 interactions were supported by text mining of 285 peer-reviewed journal publications. There were 4 interactions with Complete agreement (172 references), shown in Table 2 and 7 interactions with *Partial* agreement (113 references) reported in Table 3. The *Complete* interactions include ApoA1 → HDL particles, LDL (cholesterol) → ApoB, etc. The *Partial* interactions include cholesterol → triglycerides, etc. In the comprehensive list of NLP interactions (supplementary table 2), in addition to the 11 interactions, 3 more are marked as *Partial* interactions because the source and target nodes are various fractions of the same biomolecule. For example, trlz → vltrlp interaction shows total triglyceride rich lipoprotein is a cause of very large triglyceride rich lipoprotein. This interaction makes sense because both the source and target are part of the same biomolecule, i.e. the triglyceride rich lipoproteins and a change in the part of the whole will reflect in the whole as well. However, due to it not having support from NLP, they were omitted from interactions supported by text mined data. The complete interactions were all direct regulations whereas the all but one of the partial interactions were indirect regulations.

The undirected edges from the PC run revealed 5 interactions that were over the consensus of 50%. On comparison with NLP output, all 5 interactions were not found at all. For each of these interactions, the source and target types were sub-types of each other as well as were not matched to MedScan database analytes. These were different fractions of complex molecules such as glucose, LDL and triglycerides interacting with each other.

Table 2: Completely matching Data Mined Interactions that were recovered from text-mined search of source and target variables out of a total of 110 inferred from the data.

	source	**target**	**Source type**	**Target type**	**Interaction**	**Direction**
1	apoa1	h123	Protein	Small Molecule	DirectRegulation: estradiol ---> LPL	Unknown
2	Cldlp	apob	Small Molecule	Protein	DirectRegulation: cholesterol ---> APOB	Unknown

3	totchol	apob	Unknown	Protein	negative DirectRegulation: cholesterol --| APOB	Negative
4	totchol	nldlc	Unknown	Small Molecule	negative DirectRegulation: cholesterol --| LDL	Negative

Table 3: Partially matching Data Mined Interactions that were recovered from text-mined search of source and target variables out of a total of 110 inferred from the data.

	source	**target**	**Source type**	**Target type**	**Interaction**	**Direction**
1	Lpir	lchdlp	Clinical Parameter	Small Molecule	positive Regulation: HDL --+> insulin sensitivity	Positive
2	Lpir	trlz	Clinical Parameter	Protein	positive Regulation: triacylglycerol lipase --+> insulin sensitivity	Positive
3	trlp	ntrlc	Small Molecule	Small Molecule	negative Regulation: cholesterol ---| triacylglycerol lipase	Negative
4	ala	weight_kg	Small Molecule	Clinical Parameter	DirectRegulation: IgG ---> INS	Unknown
5	h7p	prot	Small Molecule	Protein	HDLP with both APOA1 and APOB	Unknown
6	ntrltg	ntrlc	Small Molecule	Small Molecule	negative Regulation: cholesterol ---| triacylglycerol lipase	Negative
7	vstrlp	ntrlc	Small Molecule	Small Molecule	negative Regulation: cholesterol ---| triacylglycerol lipase	Negative

Discussion

A large potential value is derived from being able to predict causality based on only sparse experimental data and hence, it is based on several assumptions. There is a need to understand the relationships that appear anew in the predicted graphs and conduct suitable experiments to confirm the hypothesis. The use of tools like pcalg for different biological data can prove to be highly useful to see the correctness of these algorithms and to be able to validate the finding of these graphs more easily. The data provided to us was a sparsely spaced data from pre- and post-interventions, just like many other datasets from the field of biology. It was interesting to see a largely continuous data prediction to get an even more thorough causal inference pathway. The PC algorithm is not efficient in such cases, hence, the use of more advanced tools like FCI (Spirtes, Meek, and Richardson 2013) which take into account the heuristics of using unknown confounders to measure the causal relationship between nodes using a Markov assumption or faithfulness assumption. For a time series data, a new approach shall be taken to discover the causal inference pathways.

The naïve PC algorithm is very reliable due to false positive rates being small as Bühlmann, Rütimann, and Kalisch (2013), tried to recover 1000 true interventions effects from 234 intervention experiments in a *S. cerevisiae* gene expression data set: by assigning 100 interventions as true and 900 as false. The algorithm was able to produce an expected logarithmic shape ROC (receiver operating characteristic) curve of true and false positives. However, it is variable position dependent, i.e. the position of variables in the input directly affect the Completed Partially Directed Acyclic Graph (CPDAG). To resolve this issue, columns were resampled for 10,000 iterations to reduce the effect of positionality of variables.

The current approaches of causal discovery pathways face certain limitations that are ubiquitous to a large number of domains similar to our data of interest, i.e. the physiological domain. One such problem rises from the missing data which could be due to random chance. In the case of a low number of observations, the output is less reliable because statistically it has more likelihood of representing a false positive result (Dumas-Mallet et al. 2017). In this case, it is hypothesized that depending on the quantity of missing data, it should not affect the outcome of a causal discovery pathway. In other cases, where the data is not missing at random, a newer discovery pathway can be recovered which is different from a causal discovery pathway

predicted from an incomplete dataset. Tu et al. (2020) discusses how the missing data causes certain edges of a graph to behave differently due to certain assumptions such as faithful observability (i.e. Edges recovered from observed data also hold true in unobserved data). Another set of assumptions include missingness indicators (representing the status of missingness) that cannot be either a deterministic cause or be used to create an edge with another missingness indicator (Tu et al. 2020). In the STRRIDE data, such a pattern was observed as well. Many of the rows have a multitude of missing data. Hence, a cutoff can be chosen to separate the highly missing data from the whole and so they are not considered for further analysis. There is also an instance where indirect relations are recovered from the data, i.e., they appear at a happenstance. Schlegel and Shapiro (Schlegel and Shapiro 2013) talk about performance of data driven methods being directly affected by incompleteness of data and how different methods can be improved by using the Wh-questioning algorithm to improve the completeness, henceforth, improving the reliability of the predictive results in cognitive systems.

The statistical tests showed only a few of the analytes were significantly changed over the two-time point study. This could be attributed to the high number of missing values in certain columns. The missing data in this study was found to be a problem in PC due high volumes of rows missing data. Especially in the **mir** columns which were micro-RNA concentrations where across the various combinations of amount and intensity groups, low amount and moderate intensity group consisted of only missing values. This created runtime errors while running the PC algorithm as it could not work by comparing missing values to themselves. The cutoff of greater than 5 NA ensured that there were no more than 5 missing values in each of the group out of the size of roughly 30 values in each group. Even with the missing values it was seen than regardless of gender, the Matsuda index (measure of insulin resistance) and blood pressures were significantly changed. This shows that participants of the study would have lowered levels of blood pressure and slightly elevated levels of Matsuda, regardless of their intervention subscription. Hence, blood pressure and insulin sensitivity are affected by an exercise regimen. The graph produced for a PC run on the pre- and post-intervention analytes was many interactions that had edges which were nonsensical as shown in supplementary figure 1. These edges did not have a biological meaning. For example, the pre intervention low density lipoprotein (ldlz_pre) appeared to be a cause of intensity of exercise and post intervention calibrated HDL concentrations (h5p_post) be a cause of pre intervention calibrated HDL

concentration (h5p_pre). Both edges cannot exist due to intensity being an independent variable than can only affect some or all analytes as well as post intervention analytes cannot affect a past intervention analyte albeit the same analyte. It also had a lot of orphan nodes along with two islands (smaller DAGs) which does not provide us with much valuable information due to fracturing of an expected single large DAG which would entail inter-relations among the islands.

A retention of edge-based strategy showed more reliable results because every edge in the final graph is at least appearing 5,000 out of 10,000 times. If interactions appeared less than 50% of time, there were instances of the same interactions, but the source and target nodes were flipped. This created bidirectionality which was talked about earlier. As was seen in the 45% retention graph, having bidirectionality in a graph that was extracted from unidirectional output of PC would make it highly unreliable. This could cause reporting of indirect relations that would appear only due to chance and have a minimal significance in discovering new relations yet add to the work done to recover true novel relations that could be derived. While the accuracy of the PC algorithm is relatively high,

The resolution of analytes used in the study was very high (multiple sized fractions of the same biomolecule). This creates a problem because the STRRIDE study used different fractions of HDL (high density lipoprotein), LDL, TG-RL (Triglyceride rich lipoproteins) which could not be fed to the NLP because it only takes in general terms such as HDL instead of a fractional term. The reason it takes HDL as an input and not a specific fraction is that almost no studies have been conducted on fractions of biomolecules or they are not published in the public journal database that comprises our knowledgebase. Hence interactions comprised of purely fractional terms have to be called *None* due to low support or *Partial* in the case of a different molecular fraction being a cause of the first one.

Interactions such as, for example, the low-density lipoprotein (LDL) is known to inhibit the cholesterol synthesis pathway (Bhanpuri et al. 2018; Wägner et al. 2002) which was observed as a cause-effect relation by our graph, hence, supporting the already existing information. Promotion/Inhibition expression was not observed in the output of PC as this would be out of scope for this study, studied in-vitro by the researchers. Examples like these provide only a level of certainty to our methodology which leaves a gap to study the undiscovered relations. To reliably discover new edges and hypotheses to test upon, we need to be able to show such interactions that already exist in nature as a sanity check for newer discoveries to be made.

Our study was able to provide literature confirmation in roughly one-sixths of the recovered relations. The remaining five-sixths of the interactions consisted of a mix of novel interactions that have not been reported as well as others that do not need to be explored as talked about earlier. While the literature-based search helped with identifying interactions that have been either well studied or those that are currently being discovered, it is certain that only a fraction of the actual number of naturally occurring interactions have been reported in MedScan. Also, the interactions that were reported by MedScan are limited to only the source and method that was used to build the tool. The natural language processing rules (used in MedScan) are also ever growing since human language and cognitive skills are vast and cannot be contained with the current methods and require further research as well (Wang et al. 2020).

The novel-like interactions found in this study shall provide grounds to research further based on consensus-based strategies that can be developed by using a multitude of algorithms who strive to discover accurate causal inference pathways. Using algorithms that use simpler mathematical modelling like ordinary differential equations and probability distributions with literature-based conformance shall prove to be provide novel inference pathways for further experimentation (Vashishtha et al. 2015; Huynh-Thu and Sanguinetti 2015). The consensus-based strategy in conjunction with data-text hydrid method provided support for validation of already existing knowledge and an insight into what can we narrow our research to be focused on to provide us with discovery of causal relations. As we research the data-text driven hybrid methods, we shall keep improving on the discovery aspect of purely statistical algorithms.

Future Work

While the data-text hybrid methods are well corroborated, since our current knowledge is used to get meaningful information out of observational data, the sources of the text-driven methods also play a key role in discovery of new causal inference interactions. There are a multitude of databases that can be utilized to this effect, including the Human Metabolome Database (https://hmdb.ca/), MetaCyc (https://metacyc.org/), etc. as well as manually curated databases. While these databases can expand the horizon of the text driven portion, we can also employ strategies to improve the false discovery rate, redundancy of similar interactions reported as separate by the data-driven algorithms, improvement of underlying causal inference method,

Using the PC algorithm, it was observed that most of the predicted graph might have around 10% false positives, which paves way for developing strategies for reducing them. The algorithm PC-p (Strobl, Spirtes, and Visweswaran 2019) seems to perform a reliable task of sorting the false positives out by evaluating p-values for edges and then ranking them to find a more accurate graph with a higher Confidence Interval. We can test the novel interactions in a laboratory experiment to validate the novel interactions to add to the regulatory and metabolomic biology. Using the same approach, we can develop graphs for other data that has not been studied well.

The consensus-based strategy can also be improved further by executing the PC algorithm even more than 10,000 runs to see if the redundant interactions disappear or fall into lower thresholds of the consensus based strategy. There are also interactions that PC algorithm might deem insignificant, however, might turn out to play key roles in certain biological areas. To find these interactions, we need to develop an even more robust underlying conditional independence methodology with the help of mathematical models used in such discoveries. This could also help us reduce the orphan graphs that are seen at higher thresholds (100% rccovery) and provide us with a wholistic picture at higher confidence levels than the studied 50% recovery graph.

The PC algorithm is one of many approaches, scientists have employed for discovery of causal inference pathways. The ODLP (Statnikov et al. 2015) algorithm is another algorithm designed to discover cause-and-effect interactions by using Markov models to find the accurate interactions of a variable T and test its faithfulness in its respective graph. Multiple approaches can be applied to a given dataset and a consensus of these approaches shall be statistically sound to make real-life decisions for developing drugs to fight virtually any lifestyle / chronic disorder. There are endless possibilities for the use case of such algorithms that need to be explored.

References

Bhanpuri, Nasir H., Sarah J. Hallberg, Paul T. Williams, Amy L. McKenzie, Kevin D. Ballard, Wayne W. Campbell, James P. McCarter, Stephen D. Phinney, and Jeff S. Volek. 2018. "Cardiovascular Disease Risk Factor Responses to a Type 2 Diabetes Care Model Including Nutritional Ketosis Induced by Sustained Carbohydrate Restriction at 1 Year: An Open Label, Non-Randomized, Controlled Study." *Cardiovascular Diabetology* 17 (1): 1–16. https://doi.org/10.1186/s12933-018-0698-8.

Bühlmann, Peter, Philipp Rütimann, and Markus Kalisch. 2013. "Controlling False Positive Selections in High-Dimensional Regression and Causal Inference." *Statistical Methods in Medical Research* 22 (5): 466–92. https://doi.org/10.1177/0962280211428371.

Cyr, Anthony R., and Frederick E. Domann. 2011. "The Redox Basis of Epigenetic Modifications: From Mechanisms to Functional Consequences." *Antioxidants and Redox Signaling* 15 (2): 551–89. https://doi.org/10.1089/ars.2010.3492.

Dumas-Mallet, Estelle, Katherine S. Button, Thomas Boraud, Francois Gonon, and Marcus R. Munafò. 2017. "Low Statistical Power in Biomedical Science: A Review of Three Human Research Domains." *Royal Society Open Science* 4 (2). https://doi.org/10.1098/rsos.160254.

Ferreiro, Diego U., Elizabeth A. Komives, and Peter G. Wolynes. 2014. "Frustration in Biomolecules." *Quarterly Reviews of Biophysics* 47 (4). https://doi.org/10.1017/S0033583514000092.

Hauser, Alain, and Peter Bühlmann. 2012. "Characterization and Greedy Learning of Interventional Markov Equivalence Classes of Directed Acyclic Graphs." *Journal of Machine Learning Research* 13: 2409–64.

Huynh-Thu, Vân Anh, and Guido Sanguinetti. 2015. "Combining Tree-Based and Dynamical Systems for the Inference of Gene Regulatory Networks." *Bioinformatics* 31 (10): 1614–22. https://doi.org/10.1093/bioinformatics/btu863.

Hyttinen, Antti, Sergey Plis, Matti Järvisalo, Frederick Eberhardt, and David Danks. 2016. "Causal Discovery from Subsampled Time Series Data by Constraint Optimization." *JMLR Workshop and Conference Proceedings* 52: 216–27. http://www.ncbi.nlm.nih.gov/pubmed/28203316%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5305170.

Johnson, Johanna L., Cris A. Slentz, Leanna M. Ross, Kim M. Huffman, and William E. Kraus. 2019. "Ten-Year Legacy Effects of Three Eight-Month Exercise Training Programs on Cardiometabolic Health Parameters." *Frontiers in Physiology* 10 (APR): 1–9. https://doi.org/10.3389/fphys.2019.00452.

Kalisch, Markus, and Peter Bühlmann. 2007. "Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm." *Journal of Machine Learning Research* 8: 613–36.

Kalisch, Markus, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. 2012. "Causal

Inference Using Graphical Models with the R Package Pcalg." *Journal of Statistical Software* 47 (11): 0–26. https://doi.org/10.18637/jss.v047.i11.

Kraus, W. E., C. E. Torgan, B. D. Duscha, J. Norris, S. A. Brown, F. R. Cobb, C. W. Bales, et al. 2001. "Studies of a Targeted Risk Reduction Intervention through Defined Exercise (STRRIDE)." *Medicine and Science in Sports and Exercise* 33 (10): 1774–84. https://doi.org/10.1097/00005768-200110000-00025.

Las Rivas, Javier de, and Celia Fontanillo. 2010. "Protein-Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks." *PLoS Computational Biology* 6 (6): 1–8. https://doi.org/10.1371/journal.pcbi.1000807.

Minde, David P., Martina Radli, Federico Forneris, Madelon M. Maurice, and Stefan G.D. Rüdiger. 2013. "Large Extent of Disorder in Adenomatous Polyposis Coli Offers a Strategy to Guard Wnt Signalling against Point Mutations." *PLoS ONE* 8 (10): 1–9. https://doi.org/10.1371/journal.pone.0077257.

Nikolay, Fabio, Marius Pesavento, George Kritikos, and Nassos Typas. 2017. "Learning Directed Acyclic Graphs from Large-Scale Genomics Data." *Eurasip Journal on Bioinformatics and Systems Biology* 2017 (1). https://doi.org/10.1186/s13637-017-0063-3.

Novichkova, Svetlana, Sergei Egorov, and Nikolai Daraselia. 2003. "MedScan, a Natural Language Processing Engine for MEDLINE Abstracts." *Bioinformatics* 19 (13): 1699–1706. https://doi.org/10.1093/bioinformatics/btg207.

Paul Shannon, 1, 1 Andrew Markiel, 2 Owen Ozier, 2 Nitin S. Baliga, 1 Jonathan T. Wang, 2 Daniel Ramage, 2 Nada Amin, 5 Benno Schwikowski, 1, 5 and Trey Ideker2, 3, 4, 山本隆久, 豊田直平, 深瀬吉邦, and 大森敏行. 1971. "Cytoscape: A Software Environment for Integrated Models." *Genome Research* 13 (22): 426. https://doi.org/10.1101/gr.1239303.metabolite.

Rathnam, Chandramouli, Sanghoon Lee, and Xia Jiang. 2017. "An Algorithm for Direct Causal Learning of Influences on Patient Outcomes." *Artificial Intelligence in Medicine* 75: 1–15. https://doi.org/10.1016/j.artmed.2016.10.003.

Salon, John a, David T Lodowski, and Krzysztof Palczewski. 2011. "The Significance of G Protein-Coupled Receptor." *Pharmacological Reviews* 63 (4): 901–37. https://doi.org/10.1124/pr.110.003350.901.

Scheines, Richard. 1997. "An Introduction to Causal Inference."

Schlegel, Daniel R, and Stuart C Shapiro. 2013. "Inference Graphs: A Roadmap." *Advances in Cognitive Systems*, no. DECEMBER 2013: 217–34.

Silverstein, Craig, Sergey Brin, Rajeev Motwani, and Jeff Ullman. 2000. "Scalable Techniques for Causal Data Mining." *Data Mining and Knowledge Discovery* 4: 163–92.

Spirtes, Peter, and Clark Glymour. 1991. "An Algorithm for Fast Recovery of Sparse Causal Graphs." *Social Science Computer Review* 9 (1): 62–72. https://doi.org/10.1177/089443939100900106.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*. 2nd ed.

Spirtes, Peter, Christopher Meek, and Thomas Richardson. 2013. "Causal Inference in the Presence of Latent Variables and Selection Bias." *ArXiv Preprint ArXiv* 1302 (4983). https://doi.org/10.7551/mitpress/2006.003.0009.

Statnikov, Alexander, Sisi Ma, Mikael Henaff, Nikita Lytkin, Efstratios Efstathiadis, Eric R. Peskin, and Constantin F. Aliferis. 2015. "Ultra-Scalable and Efficient Methods for Hybrid Observational and Experimental Local Causal Pathway Discovery." *Journal of Machine Learning Research* 16: 3219–67.

Strobl, Eric V. 2019. "A Constraint-Based Algorithm for Causal Discovery with Cycles, Latent Variables and Selection Bias." *International Journal of Data Science and Analytics* 8 (1): 33–56. https://doi.org/10.1007/s41060-018-0158-2.

Strobl, Eric V., Peter L. Spirtes, and Shyam Visweswaran. 2019. "Estimating and Controlling the False Discovery Rate of the PC Algorithm Using Edge-Specific P-Values." *ACM Transactions on Intelligent Systems and Technology* 10 (5). https://doi.org/10.1145/3351342.

Subramaniam, Shankar, Eoin Fahy, Shakti Gupta, Manish Sud, Robert W. Byrnes, Dawn Cotter, Ashok Reddy Dinasarapu, and Mano Ram Maurya. 2011. "Bioinformatics and Systems Biology of the Lipidome." *Chemical Reviews* 111 (10): 6452–90. https://doi.org/10.1021/cr200295k.

Tu, Ruibo, Cheng Zhang, Paul Ackermann, Karthika Mohan, Hedvig Kjellström, and Kun Zhang. 2020. "Causal Discovery in the Presence of Missing Data." In *AISTATS 2019 - 22nd International Conference on Artificial Intelligence and Statistics*. Vol. 89.

Vashishtha, Saurabh, Gordon Broderick, Travis J.A. Craddock, Mary Ann Fletcher, and Nancy G. Klimas. 2015. "Inferring Broad Regulatory Biology from Time Course Data: Have We Reached an Upper Bound under Constraints Typical of in Vivo Studies?" *PLoS ONE* 10 (5): 1–27. https://doi.org/10.1371/journal.pone.0127364.

Wägner, A. M., O. Jorba, M. Rigla, E. Alonso, J. Ordóñez-Llanos, and A. Pérez. 2002. "LDL-Cholesterol/Apolipoprotein B Ratio Is a Good Predictor of LDL Phenotype B in Type 2 Diabetes." *Acta Diabetologica* 39 (4): 215–20. https://doi.org/10.1007/s005920200037.

Wang, Jing, Huan Deng, Bangtao Liu, Anbin Hu, Jun Liang, Lingye Fan, Xu Zheng, Tong Wang, and Jianbo Lei. 2020. "Systematic Evaluation of Research Progress on Natural Language Processing in Medicine over the Past 20 Years: Bibliometric Study on Pubmed." *Journal of Medical Internet Research* 22 (1): 1–19. https://doi.org/10.2196/16816.

Appendix

Supplementary Table 1: A comprehensive list of analytes reported in the STRRIDE studies (From Key, missing some data)

Acac	Acetyl acetone
Acetone	Acetone
Age	Chronologic age
Ala	Alpha-linolenic acid
Albumin	Albumin (mg/dl)
Alp	Alkaline phosphatase mg/dl
Amount	Amount of exercise (low/high)
Apoa1	Apolipoprotein a1
Apob	Apolipoprotein b
Avo2	Absolute vo2 max
Bcaa	Branched-chain amino acid
Bun	Blood urea nitrogen mg/dl
Chdlp	Calibrated hdl particle
Cldlp	Calibrated ldl particle
Cmv	Cmv igg (od)
Creat	Creatinine mg/dl
Crp	Crp mg/dl
Ctr	Citrate
Dhea	Dehydroepiandrosterone
Dht	Dihydrotestosterone
Fat	Fat mass in kg pre intervention
Fat	Fat mass in kg
Gh	Growth hormone
Glu	Glucose (mg/dl)
Glut-4	Muscle glucose transporter
Glyca	Glycoprotein acetyls
Gsp	Glycated serum protein (micromoles/l)
H123	Small hdl particles <9 nm (μmol/l)
H1p	Calibrated hdl particle 7.4μmol/l
H2p	Calibrated hdl particle 7.8μmol/l
H4p	Calibrated hdl particle 8.7μmol/l
H5p	Calibrated hdl particle 9.5μmol/l
H6p	Calibrated hdl particle 10.3μmol/l
H7p	Calibrated hdl particle 10.8μmol/l
Hdlz	Hdl 7.4-13 nm
Height	Height in meters
Igf-1	Insulin gorwth factor
Infx	Inflammation index
Intensity	Intensity of exercise (moderate/vigorous)
Isi	Insulin sensitivity

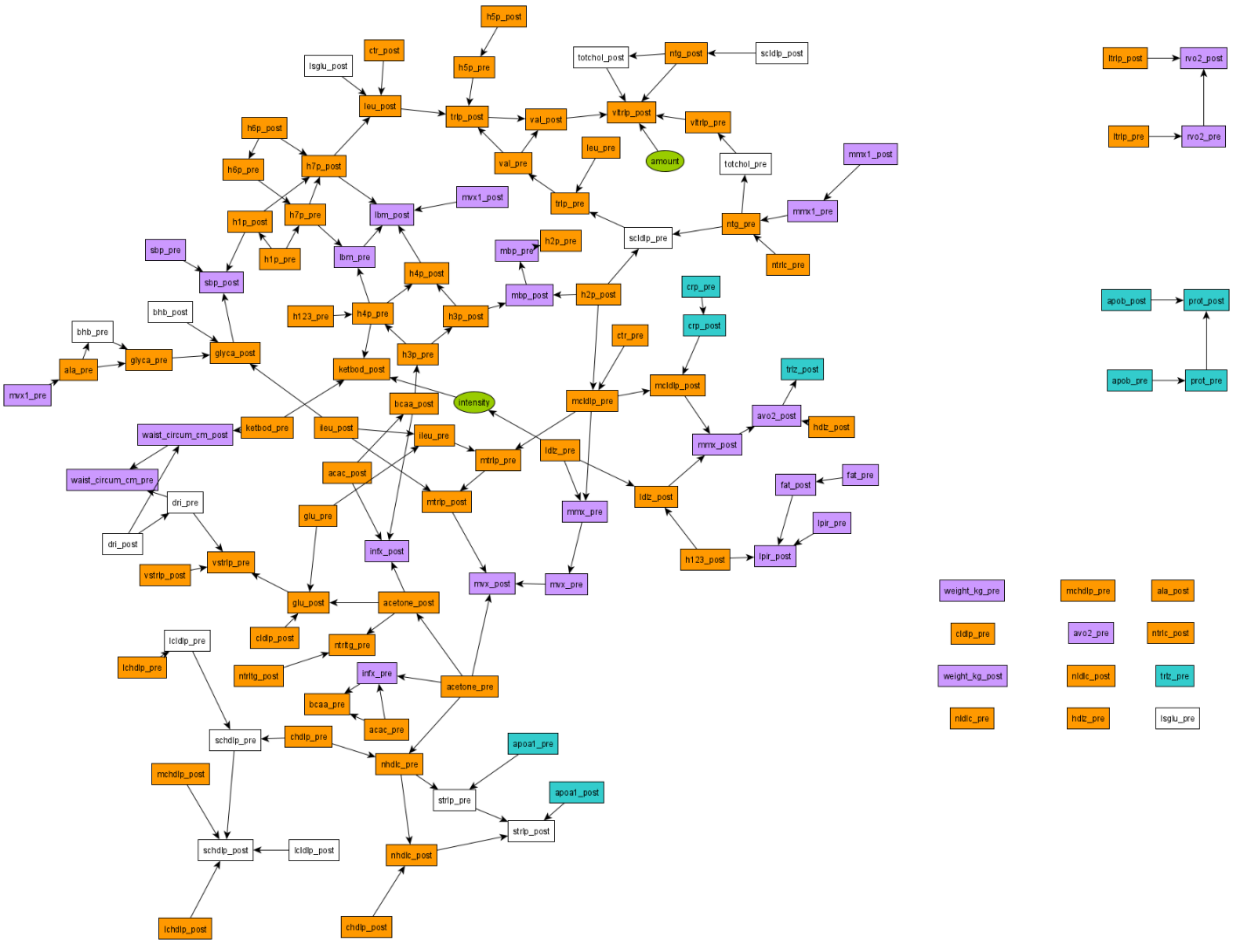
Ketbod	Ketone body
Lbm	Lean body mass in kg
Lbm	Lean body mass in kg
Lchdlp	Large chdlp 9.6-13 μmol/l
Ldlz	Mean lipoprotein sizes 19-22.5 nm
Leu	Leucine
Lpir	Lipoprotein insulin resistance index
Ltrlp	Large triglyceride-rich lipoprotein (trlp) particle
Mbp	Mean blood pressure
Mchdlp	Medium chdlp 8.1-9.5 μmol/l
Mcldlp	Medium cldlp 20.5-21.4 nmol/l
Mmx	Metabolic malnutrition index
Mmx1	Metabolic malnutrition index 1
Mtrlp	Medium triglyceride-rich lipoprotein (trlp) particle 37-49 nmol/l
Mvx	Metabolic vulnerabilty index
Mvx1	Metabolic vulnerabilty index 1
Nhdlc	Cholesterol hdl concentration
Nldlc	Cholesterol ldl concentration
Ntrlc	Colesterol fractions
Ntrltg	Trl triglyceride
Prot	Protein
Rvo2	Relative vo2 max
Sbp	Systolic blood pressure (mmhg)
Schdlp	? – unknown
Scldlp	? – unknown
Strlp	? – unknown
T3	Triiodothyronine
T4	Thyroxine
Totchol	? – unknown (Total Cholesterol?)
Trlp	Total triglyceride rich lipoprotein particles
Trlz	Mean lipoprotein sizes 30-100 nm
Tsh	Thyroid-stimulating hormone
Val	Valine
Vltrlp	Very large triglyceride-rich lipoprotein (trlp)
Vstrlp	Very small triglyceride-rich lipoprotein (trlp)
Waist_circum_cm	Waist circumference
Weight_kg	Weight in kilogram
Weight_kg	Weight in kg

Supplementary Table 2: A comprehensive list of all the interaction that were discovered through applying the PC algorithm and conformance with the Pathway Studio natural language processing of the terms from the nodes.

source	target	source_type	target_type	Data Mined	Interaction	Interaction Sub	Direction
apoa1	h123	Protein	Small Molecule	Complete	DirectRegulation: estradiol ---> LPL	DirectRegulation	unknown
cldlp	apob	Small Molecule	Protein	Complete	DirectRegulation: cholesterol ---> APOB	DirectRegulation	unknown
lpir	lchdlp	Clinical Parameter	Small Molecule	Partial	positive Regulation: HDL --+> insulin sensitivity	Regulation	positive
lpir	trlz	Clinical Parameter	Small Molecule	Partial	positive Regulation: triacylglycerol lipase --+> insulin sensitivity	Regulation	positive
totchol	apob	Unknown	Protein	Complete	negative DirectRegulation: cholesterol ---| APOB	DirectRegulation	negative
totchol	nldlc	Unknown	Small Molecule	Complete	negative DirectRegulation: cholesterol ---| LDL	DirectRegulation	negative
trlp	ntrlc	Small Molecule	Small Molecule	Partial	negative Regulation: cholesterol ---| triacylglycerol lipase	Regulation	negative
acac	ketbod	Small Molecule	Small Molecule	None			
acetone	vltrlp	Small Molecule	Small Molecule	None			
amount	rvo2	independent	Clinical Parameter	None			
avo2	rvo2	Clinical Parameter	Clinical Parameter	None	both are related to VO2max		
bcaa	mmx1	Small Molecule	Clinical Parameter	None	check type of bcaa		
chdlp	h123	Small Molecule	Small Molecule	None	both are HDL particles		
ctr	glyca	Small Molecule	Small Molecule	None			
ctr	mmx	Small Molecule	Clinical Parameter	None			
glyca	infx	Small Molecule	Clinical Parameter	None			
h123	infx	Small Molecule	Clinical Parameter	None			
h1p	acetone	Small Molecule	Small Molecule	None			
h2p	mcldlp	Small Molecule	Small Molecule	None	HDLP size vs LDLP size		
h4p	lchdlp	Small Molecule	Small Molecule	None	HDLP size vs LDLP size		
h6p	h5p	Small Molecule	Small Molecule	None	Different HDLP sizes		
h6p	hdlz	Small Molecule	Small Molecule	None	Different HDLP sizes		
hdlz	cldlp	Small Molecule	Small Molecule	None	HDLP size vs LDLP size		
hdlz	h7p	Small Molecule	Small Molecule	None	both are HDL particles		
hdlz	lchdlp	Small Molecule	Small Molecule	None	HDLP size vs LDLP size		
infx	mvx	Clinical Parameter	Clinical Parameter	None			
intensity	h1p	independent	Small Molecule	None			

intensity	rvo2	independent	Clinical Parameter	None			
ketbod	acetone	Small Molecule	Small Molecule	None			
lbm	fat	Clinical Parameter	Clinical Parameter	None	calculate weight		
lbm	weight_kg	Clinical Parameter	Clinical Parameter	None			
lchdlp	h5p	Small Molecule	Small Molecule	None	Different HDLP sizes		
ldlz	scldlp	Small Molecule	Unknown	None	Different LDLP sizes		
leu	bcaa	Small Molecule	Small Molecule	None			
leu	mmx1	Small Molecule	Clinical Parameter	None			
ltrlp	lpir	Small Molecule	Clinical Parameter	None			
mchdlp	schdlp	Small Molecule	Unknown	None	Different sizes of HDLP		
mmx	mmx1	Clinical Parameter	Clinical Parameter	None			
mmx	mvx	Clinical Parameter	Clinical Parameter	None			
mtrlp	avo2	Small Molecule	Clinical Parameter	None			
mtrlp	ctr	Small Molecule	Small Molecule	None			
mvx	mvx1	Clinical Parameter	Clinical Parameter	None			
mvx1	mmx1	Clinical Parameter	Clinical Parameter	None			
nhdlc	h123	Small Molecule	Small Molecule	None	chol conc vs HDL		
nldlc	apob	Small Molecule	Protein	None			
nldlc	cldlp	Small Molecule	Small Molecule	None	chol conc vs cLDLP		
nldlc	mcldlp	Small Molecule	Small Molecule	None	chol conc vs cLDLP		
ntrltg	mtrlp	Small Molecule	Small Molecule	None	Triglyceride rich lipoprotein vs triglyceride		
prot	ctr	Protein	Small Molecule	None			
prot	mmx1	Protein	Clinical Parameter	None			
rvo2	fat	Clinical Parameter	Clinical Parameter	None			
scldlp	cldlp	Unknown	Small Molecule	None	Different sizes of LDLP		
scldlp	mcldlp	Unknown	Small Molecule	None	Different sizes of LDLP		
strlp	trlz	Unknown	Small Molecule	None			
strlp	vstrlp	Unknown	Small Molecule	None	differetn sizes of trlp		
val	bcaa	Small Molecule	Small Molecule	None			
fat	weight_kg	Clinical Parameter	Clinical Parameter	None	fat mass contributes to weight		
ala	weight_kg	Small Molecule	Clinical Parameter	Partial	DirectRegulation: IgG ---> INS	DirectRegulation	unknown
h7p	prot	Small Molecule	Protein	Partial	HDLP with both APOA1 and APOB		
ntrltg	ntrlc	Small Molecule	Small Molecule	Partial	negative Regulation: cholesterol ---| triacylglycerol lipase	Regulation	negative

trlp	vstrlp	Small Molecule	Small Molecule	Partial	Both are lipoprotein	negative DirectRegulation: cholesterol ---| LDL	DirectRegulation
trlz	h1p	Small Molecule	Small Molecule	Partial	Both are lipoprotein	negative DirectRegulation: cholesterol ---| LDL	DirectRegulation
trlz	vltrlp	Small Molecule	Small Molecule	Partial	Both are lipoprotein	negative DirectRegulation: cholesterol ---| LDL	DirectRegulation
vstrlp	ntrlc	Small Molecule	Small Molecule	Partial	negative Regulation: cholesterol ---| triacylglycerol lipase	Regulation	negative



Supplementary Figure 1: An intial PC algorithm graph for pre vs post values that differed significantly. There are twelve orphan nodes without edges and two smaller acyclic directed graphs.

Supplementary Figure 2: A stacked barplot of missing value (NA) counts in the data selected with less than 6 NA values. The Low amount Vigorous intensity group had the maximum number of missing values whereas the least number of missing values were observed in High amount and Moderate intensity intervention groups.